UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

March 25, 1999

DSSD Census 2000 Procedures and Operations Memorandum Series # R - 2

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Subject: Census 2000 Accuracy and Coverage Evaluation Survey: Sample Allocation and Poststratification Plans

The attached paper was prepared to address the changes resulting from the January 1999 Supreme Court ruling against the use of sampling for apportionment. This paper was presented at the National Academy of Science Expert Review Panel on the 2000 Census. Consequently, the Integrated Coverage Measurement Survey (ICM) will be redesigned as an Accuracy and Coverage Evaluation Survey (ACE). Our purpose is to addresses assumptions, research and timing implications related to this redesign effort. The focus is on two key sampling and estimation issues:

- How to allocate the ACE sample?
- How to define the poststrata for estimation?

These plans are subject to change as we work our way through the research and the implications.

This report is a result of a combined effort by Census Bureau staff. Acknowledgments are due to the following persons for their timely contributions and helpful advice: Jim Farber, Robert Fay, Deborah Fenstermaker, Richard Griffin, Tom Mule, and Rajendra Singh.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List
ACE Implementation Team
Statistical Design Team Leaders

ACCURACY AND COVERAGE

EVALUATION SURVEY

PLANS FOR CENSUS 2000

Prepared for the March 19, 1999 meeting of the National Academy of Science Panel to Review the 2000 Census.

Prepared by: Donna Kostanich
Richard Griffin
Deborah Fenstermaker

March 18, 1999

ACCURACY AND COVERAGE EVALUATION SURVEY

PLANS FOR CENSUS 2000

I. INTRODUCTION

As a result of the January 1999 Supreme Court ruling against the use of sampling for apportionment, the Census Bureau is in the process of redesigning the Integrated Coverage Measurement Survey (ICM) as an Accuracy and Coverage Evaluation Survey (ACE). A primary purpose of the ICM was to produce direct state estimates with sufficient reliability for apportionment counts. The ACE is intended to evaluate the Census counts and will not be used for apportionment or even produced until after the release of the apportionment counts. This document gives preliminary plans and a research agenda for two key issues related to this survey redesign, the sample allocation and determination of the poststratification used for estimation. These plans are still in the early stages of development and are particularly subject to change as we work through the details and implications.

Background

Prior to the January 1999 Supreme Court ruling, the ICM was planned as a 750,000 housing unit sample. A key feature for the ICM design was the ability to produce direct state estimates with acceptable reliability. This feature had a direct impact on how the 750,000 sample size was allocated to each state and how the estimation poststrata were to be defined. The allocation of the ICM sample (Schindler (1998)) was designed to produce acceptable relative accuracy for each state estimate and for the apportionment result. The ICM estimation poststrata were not permitted to cross state boundaries. Thus, the challenge was to develop poststrata for each state that could be supported by the state's sample. Consequently, most states were allocated roughly equal sample sizes except for the most populous states. The primary sampling unit was a block cluster, a group of contiguous blocks with about 30 housing units. For the most part, the ICM sample was to be proportionally allocated within each state. The ICM sample design included a separate American Indian Reservation sample allocation of 350 block clusters. By the time of the Supreme Court decision, earlier commitments had become operationalized based on the ICM allocation for the 750,000 design. This imposes some constraints on the redesign effort, particularly for the allocation of the ACE sample.

The Census Bureau currently plans to develop an ACE with a reduced sample size of approximately 300,000 housing units. This is nearly twice the size of the 1990 Post-Enumeration Survey (PES). The goals are to allocate the 300,000 sample to achieve reliability for Race/Hispanic/Tenure groups across states while still attempting to maintain each state's reliability. Our plans include "borrowing strength" across States, as appropriate, during the

estimation process. State estimates will be produced by a form of synthetic estimation more similar to the 1990 PES than the direct state estimates that had been planned for the ICM. The 1990 PES synthetic estimation took into account Census region (Northeast, Midwest, South, and West), race, ethnicity, a measure of urbanicity, tenure, age and sex. This will allow the estimation process to utilize information obtained from persons in different States with similar coverage properties and thus improve the effectiveness of the State estimates, particularly for some demographic subgroups.

There are two key fundamental ACE sampling and estimation issues that need to be resolved. These are how to allocate the ACE sample and how to define the poststrata for purposes of estimation. The following sections address these issues and how we plan to proceed.

II. ACE SAMPLE ALLOCATION PLANS

Achieving the planned 300,000 housing units for the ACE by subsampling the 750,000 design for ICM is an operational necessity, but it imposes some constraints. Overall, we expect the reliability to be better than the 1990 PES for the majority of poststratum estimates. There will however be several poststrata that will be comparable or perhaps slightly less reliable than the 1990 PES, but overall we expect the state estimates to be more reliable than for the 1990 PES.

The sampling for the ACE will occur in phases:

- An initial sample of block clusters allocated according to the 750,000 ICM design.
- This initial block cluster sample will be forwarded to the field for the independent listing of housing units. (Large blocks will be sampled at a higher initial rate, allowing for later subsampling for ACE interviewing in 2000. Consequently, the actual number of housing units listed is expected to be close to 2 million.)
- The number of housing units from the independent listing will be used to determine sampling rates for the 300,000 ACE design.
- A reduction of the initial block cluster sample for ACE person interviewing. (This essentially gives the sample for estimating omissions and erroneous inclusions of persons for use in Dual System Estimation.)

In general, this plan will result in variable sampling rates within each State. The sampling rates will differ to some degree by race, ethnicity or tenure classification of the block clusters.

Time Frame

All design parameters and operational features have already been locked in for the initial sample of block clusters. At this date, Census Bureau staff are examining options for allocating the 300,000 housing units based on the 750,000 housing unit design. The time frame for developing the ACE sample allocation is limited because of operational considerations. A goal is to arrive at an acceptable allocation that is firm enough to commit to operational decisions by April, 1999. There is a critical need to know the interviewing sample sizes for planning the field staff and required resources. This date is already suspect for providing the Field adequate time to properly set up the infrastructure required for the field offices. The table below provides key ACE Sample Design Milestones. Another key activity that begins April, 1999 is the initial block cluster selection.

Key ACE Sample Design Milestones	
Sample Allocation	
April, 1999	Expected ACE sample sizes by state for field deployment
September, 1999	Specifications for reducing initial block clusters
December, 1999	Implementation of block cluster reduction
Initial Block Cluster Sample	
April, 1999	Begin initial block cluster sample selection
September, 1999	Begin field listing operation

Research Plans

The research to determine how to allocate the 300,000 housing units depends on the control variables proposed for the poststratification. The assumption is to start with the 1990 PES poststrata definitions and expand or revise as discussed in the later section. The final estimation procedure will post-stratify on age and sex as in 1990. For purposes of sample design, it is sufficient to work with the 51 poststratum groups formed by collapsing the 7 age-sex groups of the 1990 PES 357 poststratum design. See Attachment 1. The 51 poststratum groups serve as the starting point for the sample allocation. Consideration may be given to expanding the 51 groups by replacing regional poststrata for Non-Hispanic Whites with ones based on the 9 census divisions (New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific.),

The objectives for determining an ACE sample allocation are:

- Meet the 1990 level of reliability in terms of expected Coefficient of Variation (CV) for the 51 poststratum groups.

- Provide acceptable CVs for the state total population estimates.

- Reduce the differences among CVs within race/ethnicity and tenure groups.

There will be several poststrata that have expected reliability less than 1990. This can not be avoided since the 750,000 sample allocation was not designed for ACE.

Preliminary Findings

Three sample allocation alternatives have been developed for a sample size of 300,000 housing units. The first sample allocation alternative calls for a differential reduction by race and Hispanic ethnicity within states. The goal of a differential reduction is to get the sample size needed to meet survey objectives for non-White poststrata. This allocation does the best for non-White poststratum CVs. We view this as the most promising of the three and a starting point for further refinement. The second sample allocation is a proportional reduction of the 750,000 housing unit design across all states. This allocation is a benchmark and gives a quick sense of the poststratum reliability for a sample design contingent on the ICM allocation. A notable effect is the poor performance in the Midwest Black and Hispanic poststrata. The third sample alternative allocates a fixed sample size to states with small population and proportionally allocates the balance of the sample to the more populous states. The third alternative did not do quite as well for non-White poststratum CVs as the differential reduction, but for the most part, the differences are not big. For all three allocation plans, the American Indian Reservation sample will not be reduced.

The differential reduction is viewed as the most promising allocation and we will continue to refine this alternative. Since the third alternative was not much different, we will explore some variations of this type of design if time and resources permit.

For reference, Attachment 2 shows the resulting 1990 poststratum CVs. The expected poststratum CVs for each of the three allocations are given in Attachments 3, 4, and 5. Attachment 6 displays the expected state CVs based on borrowing strength for the three allocations. Below is a general comparison of the expected reliability for the differential reduction shown in Attachment 3 with the 1990 PES results.

- Even though the 2000 ACE will be almost twice the size of the 1990 PES, we do not expect that all poststratum groups will have better reliability than achieved in 1990. This is particularly evident in the Midwest poststrata. We are examining the degree to which we can control this difference through subsampling rates.

- Overall, we expect the reliability to be better than the 1990 PES for the majority of poststratum estimates.

- For regional estimates of Non-Hispanic Whites, we expect the majority of the poststratum CVs to range from 0.2 percent to 1.5 percent. The similar range for 1990 was 0.4 percent to 1.9 percent. The Non-Urban Renters in the Northeast will be somewhat higher than the other poststrata, but this is consistent with the 1990 results.

- For estimates of Blacks, we expect the majority of the poststratum CVs to range from 0.7 percent to 2.0 percent. The similar range for 1990 was 0.9 percent to 3.0 percent. The Non-Urban Renters will be somewhat higher than the other poststrata, but this is consistent with the 1990 results.

- For estimates of Hispanics, we expect the majority of the poststratum CVs to be comparable with the 1990 Non-Black Hispanic CVs which ranged from 1.0 percent to 6.1 percent.

- For estimates of Asians, we expect the CVs to be roughly half the 1990 CVs.

- For American Indians on Reservations, we expect the CV to be comparable with 1990 or better. (See the Limitations section below for more discussion of this poststratum.)

- For state total population estimates, we expect most CVs will be less than 0.5 when borrowing strength across states.

Limitations

The CVs presented are for comparing the relative accuracy of design alternatives. There is no commitment for achieving these CVs. The expected CVs are calculated by adjusting the 1990 CV to account for sample size and weight variation. Our expectations are dependent on attaining similar census and ACE results as 1990. The extent that the 2000 ACE is different from the 1990 PES will impact the reliability of estimates. Some notable differences between 1990 and 2000 include the surrounding block search method and the treatment of movers.

The basis of this allocation work is the 1990 PES data. This assumes that certain census and PES results from 1990 occur again in 2000. Some of these results include:

- The Master Address File is 99 percent complete.

- The ACE response rate is comparable to the 1990 PES response rate.

- There are similar undercount rates as in 1990.

- There is 100 percent data capture in Census 2000.

If any of the above conditions are not met then the reliability estimates provided will differ.

There are changes between the 1990 and 2000 methodologies that we attempted to reflect in the reliability estimates. These changes are:

- Sample Design: This research involves removing the differential weighting of the 1990 PES design and replacing it with the expected weighting of the allocation plan.

- Surrounding block search: In 1990, a surrounding block search of 1 to 2 rings was performed for all sampled blocks. In 2000, the current plan is to only perform a surrounding block search for a limited number of the sampled blocks. An adjustment has been made to compensate for the decrease in surrounding block search.

- Separate Asian and Hawaiian and Pacific Islander Estimates: In 1990, Asians and Pacific Islanders were combined during poststratification. For 2000, separate estimates for Asians, and Hawaiians and Pacific Islanders are being considered. This research uses the combined group design effect from 1990 to generate CVs.

There are differences between the 1990 and 2000 methodologies that have not been reflected in these estimates. These differences are:

- Handling movers: There is a time lag between Census day and the ACE interview day. During that period, people can and do move. The current plan uses a different procedure for handling movers as compared to 1990. This research does not reflect the difference in handling movers procedures because it is not known how the current plan for movers will affect the reliability estimates.

- American Indian Country: In 1990, the American Indian Reservation poststratum was defined to include only American Indians living on a reservation. For 2000, the poststratum definition will be expanded to include American Indians living on American Indian and Alaska Native areas (AIANAs)[1]. The types of AIANAs are American Indian reservations and trustlands, tribal jurisdiction statistical areas, tribal designated statistical areas, and Alaska Native Regional Corporations. At this point, our research does not reflect this poststratum definition change.

---

[1] For ACE, the American Indian Reservation sampling stratum is limited to people living on American Indian reservations or trustlands. For estimation, the poststratum includes all American Indians living in AIANAs.

- American Indians Off Country: In 1990, the American Indians living off reservations were included in the Non-Hispanic White and Other poststratum. For this research, all American Indians living off reservations are still grouped with the Non-Hispanic White and Others.

III. ACE POSTSTRATIFICATION PLANS

Poststratification occurs during the Accuracy and Coverage Evaluation (ACE) estimation phase and involves grouping sample elements (persons) into approximately "homogeneous" groups with respect to census coverage.

The purpose of poststratification is to reduce bias in the coverage estimates (e.g., correlation bias) without too much of an increase in the variance of these estimates.

Issues that affect poststratification of estimation elements (persons) include:

- A requirement for a minimum sample size to control sampling variance of estimates and ratio estimation bias

- Differences in reporting for the census enumeration and ACE resulting in different poststrata classifications for the same person for Dual System Estimation (DSE).

- Defining poststrata such that persons within the same poststratum have similar capture probabilities (limit heterogeneity).

- Collapsing of poststrata (defined to reduce heterogeneity) due to the minimum sample size requirement. Collapsing reduces the variance of DSEs but it increases the heterogeneity. Collapsing methodology must take into account both variance and bias.

BACKGROUND

For the 1990 Post Enumeration Survey (PES) 51 major poststrata defined by race/origin, tenure, urbanization and region were formed (See Attachment 1). Within each of these 51 groups, 7 age/sex poststrata were formed resulting in 357 final poststrata. The goal was to address the homogeneity assumption required for unbiased estimation using the DSE. This assumption requires that all people in a poststratum have the same probability of being counted in the census. We know this is not true for the total population. After completion of the 1990 PES, research

provided evidence of residual heterogeneity in the 357 poststrata design. Alternatives building upon Hard-to-Count (HTC) scores and predicted inclusion probabilities were investigated. The appendix provides details of the development of the 357 poststrata design as well as the research on these alternatives.

RESEARCH ON POSTSTRATIFICATION VARIABLES FOR CENSUS 2000

Poststratification research for the Census 2000 ACE will start with the 357 poststrata design from the 1990 PES and expand and/or refine the design since we are planning to have twice the sample size and developments since 1990 indicate it may be possible to improve the 357 poststrata design. These 357 poststrata are comprised of 7 age/sex groups cross-classified by the 51 major poststrata. The major poststrata, illustrated in Attachment 1, are defined by race/origin, tenure, urbanization and region. There is still some question as to the importance of the poststratification variables used for the 1990 PES 357 poststrata design; we know that some heterogeneity remains in these poststrata. Based on previous research, experience, policy considerations, and anticipated ACE sample allocation, we would expect to:

- Expand the 5 race/ethnicity groups to 6 by creating separate Asian poststrata and Native Hawaiian or Pacific Islander poststrata.

- Retain the regional and tenure variables.

- Replace regional poststrata for Non-Hispanic Whites with ones based on the 9 census divisions (New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific.)

- Replace the urbanicity variable. Urban/rural definitions will not be available in time for production poststratification for Census 2000 ACE. Thus, if urbanicity is determined to be an important variable we will attempt to find a suitable proxy variable.

Other promising variables, such as mail return rates, are being considered and are identified in the next section which describes the research methodology. It is always possible that the research could reveal variables more important than those listed above.

Research Methodology

Research is being performed using logistic regression techniques to assess the importance of each variable from the 357 poststrata design as well as additional variables we think may be useful. The dependent variable in our logistic regression models is a dichotomous variable for capture in the census (1 if a person is captured, 0 otherwise). Independent variables are those likely to be good predictors of the probability of a person being captured in the census.

When a logistic regression model is developed (i.e., fitting the model), the output is an estimated coefficient for each significant independent variable and its standard error. These estimated coefficients in logistic regression have an interpretation directly related to an **odds ratio**. This interpretability of the coefficients is the fundamental reason why logistic regression has proven to be a powerful analytic tool which will work remarkably well for determination of variables important in predicting probability of capture in the census. The following example explains this.

Consider the independent variable tenure (owner/renter) and the dependent census capture variable (capture/not capture). For 100 persons, the cross-classification of tenure with census capture is as follows:

	Owner	Renter	Total
Census Capture			
yes	21	22	43
no	6	51	57
Total	27	73	100

Suppose a logistic regression model is fit to predict capture in the census given a person's tenure. The **odds** of capture in the census for an owner is the ratio of the probability of an owner being captured (21/27) to the probability of an owner not being captured (6/27). This ratio is 21/6. The odds of capture in the census for a renter is similarly (22/73)/(51/73) = 22/51. The **odds ratio** for tenure is defined as the ratio of the odds for owners to the odds for renters. In this example the odds ratio is (21x51)/(6x22) = 8.11. It approximates how much more likely (or unlikely) it is for an owner to be captured in the census than a renter. When a logistic regression model is fit to the data in the above table, the estimated coefficient for tenure provides the odds ratio(the natural logarithm of the odds ratio is the estimated coefficient). Without worrying about the mathematical details, this means that the results of fitting the logistic regression model gives an estimate of the odds ratio of an independent variable. The closer the odds ratio is to 1, the less important the independent variable is in determining capture probability. An odds ratio of 1 would means an owner is just as likely to be captured as a renter and would indicate that tenure is not important in predicting census capture. An odds ratio of 2 has the same interpretation as an odds ratio of 1/2, since both indicate twice as likely for a person with one value of the predictor being captured as a person with the opposite value. In this example, an odds ratio of about 8 indicates owners are much more likely to be captured in the census and that tenure is probably an important independent variable.

If the independent variable has more than two levels, the interpretation is similar using logistic regression output to provide odds ratios of independent variables which indicate how sensitive probability of capture in the census is to varying levels of the independent variable. In our

research logistic regression models with several independent variables will be fit. The output will provide us with estimates of odds ratios as well as confidence intervals for these estimates. A comparison of these estimated odds ratios will indicate important predictors of census capture as well as a hierarchy of importance for collapsing poststrata.

The shorter Census 2000 short form has eliminated potentially interesting stratification variables from 1990 such as rent/value, type of structure, or number of rooms. The use of tract level mail return rates has been considered in previous work without much success. Adding geography as a poststratification variable has the potential to decrease the heterogeneity bias of estimates. Unfortunately geographic poststratification will increase the standard error of all small area estimates contained within a single geographic area. Demographers have questioned the use of geography as a poststratification variable noting that reducing the amount of geographic poststratification would permit other, perhaps more useful, variables to be included in defining poststrata. Research has also demonstrated that combining ACE results with Demographic Analysis (DA) may improve the overall results. However, the issue of which combining model is appropriate requires significant research and it has been decided not to combine ACE and DA for Census 2000. Research on combining will continue, however.

The following independent variables will be used in our logistic regression modeling:

- Census Region
- Census Division
- Race/Hispanic origin: (1) Non-Hispanic White & Other, (2) Black, (3) Non-Black Hispanic, (4) Asian & Pacific Islander, and (5) American Indians on Reservations.

  Note: Asian & Pacific Islander will be separated for Census 2000 ACE. They are combined for research purposes because the 1990 data files already have these races combined.
- Age/sex: (1) under 18, (2) 18 - 29 male, (3) 18 - 29 female, (4) 30 - 49 male, (5) 30 - 49 female, (6) 50 + male, and (7) 50 + female.
- Tenure: (1) owner and (2) renter.
- Type of Enumeration Area (TEA): (1) Tape Address Register (TAR) and (2) Prelist Pocket, Update Leave, and List/Enumerate
- Urbanicity: (1) urbanized areas > 250,000 , (2) other urban and (3) non-urban areas.

Note: Urban/rural definitions will not be available in time for production poststratification for Census 2000 ACE. Thus if urbanicity is determined to be an important variable a revised definition will be needed.

- Percent owner: (1) low and (2) other. Percent owner is a block-level variable. Low percent owner blocks are those blocks in the bottom 25th percentile based on percent owners.

- Mail response rate: (1) low and (2) other. Mail a response rate is a block-level variable defined as the proportion of households in the 1990 mail universe which completed their census form without the aid of an enumerator. Low mail response rate blocks are those in the bottom 25th percentile based on a mail response rate.

- Percent minority: (1) high and (2) other. Percent minority is a block-level variable. High percent minority blocks are those in the top 75th percentile based on percent minorities.

- Household size: (1) one and (2) two or more.

We will also look at the interaction of significant main effect variables. If the interaction is significant, we may be able to take it into account during collapsing.

First we will compare a hierarchy for the 357 poststrata design based on odds ratios to the hierarchy used for this design in 1990. The 1990 hierarchy was based on measures of homogeneity using measure of census performance while the logistic regression/odds ratio approach is based on how change in level for a variable effects capture probability. It is not clear what to do if the results indicate a different hierarchy. We will simulate more than one poststratification plan, if necessary, based on a different hierarchy of the 1990 PES poststratification variables.

RESEARCH ON COLLAPSING

Once we determine which variables are important and their hierarchy (possibly multiple orderings will be looked at), we need to look at collapsing. Collapsing will initially be performed as in the 1990 PES on the important variables ignoring age/sex. Once this is done we will form 7 age/sex poststrata as in 1990 within each collapsed poststrata. The important variables other than age/sex (referred to as major variables or major poststrata from now on) in 1990 were 5 race/origin categories, owner/renter, 3 urbanization categories, and the 4 Census regions. A complete cross of these variables produces 120 (5x2x3x4) potential major poststrata groups. American Indians on reservations were collapsed to one major poststratum group producing 7 poststrata by age/sex categories. We will add to this mix the most important major variables that come out of the logistic regression. This may be done several times using different additional variables. Thus, prior to collapsing we will have several different poststrata designs

defined by major variables. For each design we will have a collapsing hierarchy and we may look at several of these for a given design. Collapsing of major poststrata will occur when any major poststrata is projected based on the Census 2000 ACE sample design to have less than X (to be determined) ACE sample persons. Collapsing is done to decrease a sampling variance but it does increase bias. We want to collapse so as to improve the mean square error (variance + bias squared). If we need to collapse, the hierarchy will indicate the order of collapsing.

For each collapsed major poststrata design, we will form poststrata using the 7 age/sex categories from the 1990 PES (0-17, 18-29 M, 18-29 F, 30-49 M, 30-49 F, 50+ M, 50+ F). For example if we added two levels of mail response rate to the 1990 PES variables, we would start with 240 major poststrata prior to collapsing and then collapse based on a given hierarchy of the major variables. For any resulting poststratification we will compute coverage factors and their standard errors using Jackknife replication variance estimation methodology using the 1990 PES data. Recall that the plan for Census 2000 is for a 300,000 housing unit ACE sample while the 1990 PES sample had 150,000 housing units. We will do an adjustment of the variances to account for the increased sample size and a different allocation of sample to sampling strata. Alternative poststratification designs will be evaluated based on resulting poststrata estimated coverage factor variances.

The process described above effectively treats age/sex as the most important variables as they are never collapsed. Based on results from the logistic regression modeling, we will look at alternatives that collapse age/sex so that some variables that may be more useful in decreasing heterogeneity can be utilized.

SMALL AREA ESTIMATION

In addition, for each poststratification design, we will compute simple synthetic estimates and coverage factors for total population in each 1990 PES block cluster. For each poststratum the simple synthetic estimate is the census count in that poststratum for the block cluster multiplied by the estimated poststratum coverage factor. The total population synthetic estimate is the sum of these synthetic estimates over all poststrata. Variances for these estimates will be calculated using Jackknife methodology. Also for each poststratification design the 1990 PES block cluster estimates will be compared to a target estimate. This target will be constructed by combining all E and P sample people but removing the matches and erroneous enumerations. Thus, mean square errors can be estimated and averaged over the block clusters. Note that this target has the limitation of not including persons missed by both the census and PES. All these results will be compared for alternative poststratification methods to aid in the selection of the best one.

REFERENCES

Alho, J. M., Mulry, M. H., Wurdeman, K., and Kim, J. (1993), "Estimating Heterogeneity in the Probabilities of Enumeration for Dual System Estimation", Journal of the American Statistical Association, 88, 1130-1136.

Bell, W.R. (1993), "Using Information from Demographic Analysis in Post Enumeration Survey Estimation, Journal of the American Statistical Association, 88, 1106-1118.

Hogan, H. (1992), "The 1990 Post-Enumeration Survey: an Overview", The American Statistician, 46, 261-269.

Hogan, H. (1993), "The 1990 Post Enumeration Survey: Operations and Results", Journal of the American Statistical Association, 88, 1047-1060.

Killion, R.A. (1998), "Estimation Decisions for the Integrated Coverage Measurement Survey for Census 2000", Census 2000 Decision Memorandum No. 42.

Petroni, R., Kearney A., and Robinson J.G. (1997 A), Use of Hard-to-Count Scores and Inclusion Probabilities to Improve Dual System Estimation and Census Plus Estimates", American Statistical Association Proceedings of the Section on Survey Research.

Petroni, R. (1997 B), "Further Evaluation of the Use of Hard-to-Count Scores and Inclusion Probabilities in Post-Stratification to Improve DSE Estimates", Census 2000 Dress Rehearsal Memorandum A-13.

Schindler, E. (1998) "Allocation of the ICM Sample to the States for Census 2000," *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

Robinson J.G. (1996), "Demographic Review of the Housing and Population Results of the 1995 Test Census, Population Division's Review of the 1995 Census Test Results, Memorandum No. 2, 3/12/96, Bureau of the Census.

Thompson, J. (1992), "CAPE Processing Results", Census Bureau Memorandum, 3/20/92.

APPENDIX

1990 PES Design

For the 1990 Post-Enumeration Survey (PES) poststrata were formed to address the homogeneity assumption required for unbiased estimation using the DSE. This assumption requires that all people in a poststrata have the same probability of being counted in the census. We know this is not true for the total population. Demographic analysis estimates show that the Census has a persistent pattern of differential undercount by race, age, and sex. We also know that the difficulty of taking a Census, and the kinds of errors, differ for central cities, suburban areas, small towns, and rural areas. Previous research had also shown that minority renters are especially difficult to count. In addition, there was a belief that people living in different areas of the country might have different inclusion probabilities not easily reflected in other variables. Poststrata were formed to classify persons into groups that were as much alike as possible with respect to their Census inclusion probability. There were 1,392 poststrata formed by defining 116 groups using Census Region and Division, race/origin, place size, and tenure and then forming 12 age/sex groups within each of these 116 groups. It was anticipated that many of the 1,392 poststrata estimates would have coefficients of variation too high to be useful in the possible adjustment of the 1990 Census. Thus, a regression approach was adopted whereby a regression equation was used to predict the adjustment factor (ratio of the DSE to the census count) for each poststratum. The regression predicted factor was then combined with the direct estimate factor to form a "smoothed" factor (Hogan (1992)).

Using the results from DSE in the 1,392 poststrata the Secretary of Commerce decided not to use the PES estimates to adjust the 1990 Census. There were several criticisms of the smoothed adjustment factors produced for the 1,392 poststrata: (1) The use of smoothing models led to estimates whose true uncertainty was difficult to access; (2) The poststrata were possibly too heterogenous, especially geographically, to be suitable for the synthetic estimation of undercount for small areas; and (3) The direct (unsmoothed) estimates were thought to be biased (Hogan (1993)).

In order to respond to these criticisms, new poststrata were designed to increase homogeneity while at the same time reduce the variance of DSEs by forming fewer poststrata. In forming poststrata, one is faced with two opposing goals. First, one would like each of the poststrata to be as homogenous as possible. This can be accomplished by forming many, relatively small poststrata. But in general, for any fixed overall sample size, more poststrata means smaller sample sizes within each and thus higher variance for each of the poststrata. Because the goal was to develop fewer as well as more homogeneous poststrata, it was important to choose the stratification variables wisely. The original 116 poststrata groups had been based on a hierarchy in which geographic differences were largely maintained (i.e., combine race groups before combining geographic groups to reduce the number of poststrata) over race and ethnic differences. Differences in place/size were maintained over differences in housing tenure. The results of the PES did not necessarily validate this hierarchy. For example, differences between

some place/size categories were often very small, whereas differences between owners and renters were often striking. In developing new poststrata there were limits to using PES results directly. Thus, the analysis focused on measures of census performance derived from the complete census file, such as mail return rates and whole person substitution rates. Measures of crowding, proportion of nonhousehold members, item imputation rates and a few other variables proved helpful. The working assumption was that poststrata defined to be relatively homogeneous with respect to these variables would be relatively homogeneous with respect to the undercount. The results of this analysis suggested a hierarchy (from most important to least) of race/origin; tenure; urbanization; and region (Hogan(1992)). This hierarchy was used to collapse potential poststrata when minimal sample size constraints were violated. For example, for Blacks in Other Urban and Non-Urban areas, region was collapsed so the resulting poststrata were defined by race, urbanization, and tenure only. All Black owners in Non-Urban areas in the entire country were in the same poststrata (Thompson (1992)).

Considerable research went into deciding whether there was a grouping of states that was better than the 4 Census regions. Although some alternative patterns did emerge, none were consistent across the variables of interest (i.e., mail back rate, allocation rate, and so forth). The decision was made to use the traditional 4 Census regions because of their familiarity to users of Census products, but to drop finer breakdowns by divisions (Hogan (1993)).

The result was 51 major poststrata groups which were each divided into 7 age/sex categories producing 357 final poststrata (See Attachment 1). In the 1990 PES, data from 150,000 housing units were used to estimate undercount rates for this 357 poststrata design for the postcensal estimates. No smoothing of adjustment factors was done.

Research efforts in the 1990's

Alho, et al. (1993) and Robinson (1996) provide evidence of residual heterogeneity after implementation of the 357 poststrata design. Additionally, Bell (1993) noted that in the 1990 PES for some poststrata the Census Bureau obtained negative estimates of the number of persons missed by both the Census and the PES. Theoretically this can occur because of sampling errors. It may also occur if the data reported by Census and the PES interview differ, resulting in differing poststratification classifications for the two sources used in dual system estimation. If the variables used to form poststrata have low response variance, this negative estimate concern could be reduced. Petroni et.al. (1997 A and B) used results from the Oakland 1995 Test Census to evaluate alternative poststratification methods compared with production 1995 poststrata defined by a cross-classification of race/origin, tenure, age and sex. The goal was to produce poststrata with reduced heterogeneity and fewer negative estimates of persons missed by the initial phase enumeration and PES. The alternatives build upon the Hard-to-Count (HTC) score and inclusion probability concepts of Robinson and Alho et.al. The HTC scores are based on twelve 1990 Census variables related to undercount. The variables encompass housing conditions and population/socioeconomic dimensions. The inclusion probabilities are estimated

from logistic regression models using explanatory variables correlated with inclusion in the Census. The research formed poststrata based on:

- ranges of tract, and block group based composite HTC scores
- ranges of tract or block group based composite HTC scores crossed by various demographic characteristics such as tenure, race/origin, and age/sex.
- ranges of inclusion probabilities
- ranges of HTC scores crossed by ranges of inclusion probabilities

Research results showed that none of the alternative poststratifications examined provided improved estimates of persons missed by both the initial enumeration and the PES enumeration. Residual heterogeneity bias was found to be present in the production estimates and in the estimates produced using each of the alternatives. It was recommended that if HTC scores could be modified in a way that could reduce heterogeneity bias among owners, then the use of HTC scores for poststratification could be worth pursuing.

Hard to Count (HTC) scores will not be used for poststratification for Census 2000. This decision rests on the desire to only use the most recent data, Census 2000 data, to form groups of individuals (poststrata) believed to have similar coverage characteristics (Killion (1998)). The HTC scores have been developed for each 1990 Census tract using 1990 Census 100% and long form data. We do not want to use 1990 data to form poststrata for Census 2000. Timing makes it impossible to use Census 2000 long form data to compute HTC scores that could be used to form poststrata.

PES - 51 Major Poststrata

Non-Hispanic White & Other Poststrata

	Owner	NonOwner
Urbanized Areas	NE, MW, S, W	NE, MW, S, W
Other Urban	NE, MW, S, W	NE, MW, S, W
NonUrban	NE, MW, S, W	NE, MW, S, W

Black Poststrata

	Owner	NonOwner
Urbanized Areas	NE, MW, S, W	NE, MW, S, W
Other Urban		
NonUrban		

Non-Black Hispanic Poststrata

	Owner	NonOwner
Urbanized Areas	NE, MW, S, W	NE, MW, S, W
Other Urban		
NonUrban		

Asian & Pacific Islander Poststrata

Owner	NonOwner

American Indians on Reservations Poststratum

Date: March 15, 1999 Draft

Attachment 2

1990 PES Results
CVs for 51 Major Poststrata

Post-Strata Groups	North East	South	Midwest	West	Total
Non-Hispanic White & Other					
Owner					
Urbanized Areas 250,000 +	1.06%	0.71%	0.39%	0.65%	
Other Urban	0.49%	0.42%	0.40%	0.58%	
Non-Urban	0.69%	0.69%	1.17%	0.69%	
Non-owner					
Urbanized Areas 250,000 +	1.41%	1.52%	1.66%	1.67%	
Other Urban	1.56%	1.80%	1.11%	1.40%	
Non-Urban	4.37%	1.81%	1.56%	1.93%	
Black					
Owner					
Urbanized Areas 250,000 +	1.93%	0.92%	0.87%	2.03%	
Other Urban					1.00%
Non-Urban					1.96%
Non-owner					
Urbanized Areas 250,000 +	1.76%	2.04%	1.79%	3.02%	
Other Urban					1.23%
Non-Urban					5.68%
Non-Black Hispanic					
Owner					
Urbanized Areas 250,000 +	4.41%	0.92%	2.48%	0.90%	
Other Urban					1.68%
Non-Urban					2.75%
Non-owner					
Urbanized Areas 250,000 +	3.77%	2.82%	3.49%	1.96%	
Other Urban					2.90%
Non-Urban					6.09%
Asian & Pacific Islander					
Owner					1.48%
Non-Owner					2.70%
American Indians on Reservations					5.25%

Date: March 15, 1999 Draft Attachment 3

Differential Reduction by Race/Ethnicity Alternative
Poststrata CVs: 300,000 Housing Units

Post-Strata Groups	North East	South	Midwest	West	Total
Non-Hispanic White & Other					
Owner					
Urbanized Areas 250,000 +	0.7%	0.5%	0.2%	0.3%	
Other Urban	0.2%	0.4%	0.3%	0.4%	
Non-Urban	0.4%	0.6%	0.9%	0.5%	
Non-owner					
Urbanized Areas 250,000 +	0.8%	1.0%	1.1%	0.8%	
Other Urban	1.1%	1.1%	0.5%	1.0%	
Non-Urban	3.4%	1.5%	1.1%	1.5%	
Black					
Owner					
Urbanized Areas 250,000 +	1.9%	0.9%	1.0%	0.8%	
Other Urban					0.7%
Non-Urban					1.6%
Non-owner					
Urbanized Areas 250,000 +	1.3%	1.7%	1.8%	1.6%	
Other Urban					0.6%
Non-Urban					6.1%
Non-Black Hispanic					
Owner					
Urbanized Areas 250,000 +	3.4%	0.5%	2.1%	0.4%	
Other Urban					1.3%
Non-Urban					1.5%
Non-owner					
Urbanized Areas 250,000 +	2.1%	1.0%	4.4%	1.5%	
Other Urban					2.3%
Non-Urban					5.9%
Asian Owner					0.6%
Pacific Islander Owner					2.3%
Asian Non-owner					1.4%
Pacific Islander Non-owner					4.6%
American Indians on Reservations					3.2%

Proportional Reduction Alternative
Poststrata CVs: 300,000 Housing Units

Post-Strata Groups	North East	South	Midwest	West	Total
Non-Hispanic White & Other					
Owner					
Urbanized Areas 250,000 +	0.8%	0.5%	0.3%	0.3%	
Other Urban	0.2%	0.3%	0.4%	0.4%	
Non-Urban	0.4%	0.5%	1.2%	0.5%	
Non-owner					
Urbanized Areas 250,000 +	0.8%	0.9%	1.4%	0.8%	
Other Urban	1.2%	1.0%	0.6%	1.0%	
Non-Urban	3.6%	1.3%	1.3%	1.4%	
Black					
Owner					
Urbanized Areas 250,000 +	2.7%	1.1%	1.6%	0.7%	
Other Urban					0.7%
Non-Urban					1.7%
Non-owner					
Urbanized Areas 250,000 +	1.8%	1.9%	2.7%	1.4%	
Other Urban					0.7%
Non-Urban					6.6%
Non-Black Hispanic					
Owner					
Urbanized Areas 250,000 +	4.1%	0.6%	3.1%	0.4%	
Other Urban					1.3%
Non-Urban					1.4%
Non-owner					
Urbanized Areas 250,000 +	2.8%	1.1%	6.6%	1.3%	
Other Urban					2.3%
Non-Urban					5.8%
Asian Owner					0.6%
Pacific Islander Owner					2.4%
Asian Non-owner					1.5%
Pacific Islander Non-owner					4.6%
American Indians on Reservations					3.2%

Proportional Allocation to States Alternative
Poststrata CVs: 300,000 Housing Units

Post-Strata Groups	North East	South	Midwest	West	Total
Non-Hispanic White & Other					
Owner					
Urbanized Areas 250,000 +	0.7%	0.4%	0.2%	0.3%	
Other Urban	0.2%	0.3%	0.3%	0.4%	
Non-Urban	0.3%	0.5%	0.9%	0.5%	
Non-owner					
Urbanized Areas 250,000 +	0.7%	0.8%	1.0%	0.8%	
Other Urban	1.0%	0.9%	0.5%	1.0%	
Non-Urban	3.0%	1.2%	1.0%	1.4%	
Black					
Owner					
Urbanized Areas 250,000 +	2.3%	1.0%	1.1%	0.7%	
Other Urban					0.7%
Non-Urban					1.6%
Non-owner					
Urbanized Areas 250,000 +	1.6%	1.8%	1.9%	1.5%	
Other Urban					0.6%
Non-Urban					6.3%
Non-Black Hispanic					
Owner					
Urbanized Areas 250,000 +	3.5%	0.5%	2.1%	0.4%	
Other Urban					1.3%
Non-Urban					1.4%
Non-owner					
Urbanized Areas 250,000 +	2.5%	1.0%	4.5%	1.4%	
Other Urban					2.3%
Non-Urban					5.8%
Asian Owner					0.6%
Pacific Islander Owner					2.4%
Asian Non-owner					1.4%
Pacific Islander Non-owner					4.7%
American Indians on Reservations					3.2%

State Total Population Synthetic CVs

Region	State	1990 PES	Differential Reduction	Proportional Reduction	Proportional Allocation to States
North East	CT	0.4%	0.3%	0.3%	0.3%
North East	MA	0.5%	0.3%	0.3%	0.3%
North East	ME	0.6%	0.4%	0.4%	0.3%
North East	NH	0.6%	0.4%	0.4%	0.3%
North East	NJ	0.6%	0.4%	0.5%	0.4%
North East	NY	0.6%	0.4%	0.5%	0.4%
North East	PA	0.5%	0.3%	0.4%	0.3%
North East	RI	0.6%	0.4%	0.4%	0.3%
North East	VT	0.8%	0.5%	0.6%	0.5%
Midwest	IA	0.4%	0.3%	0.4%	0.3%
Midwest	IL	0.4%	0.3%	0.5%	0.3%
Midwest	IN	0.4%	0.3%	0.4%	0.3%
Midwest	KS	0.4%	0.3%	0.3%	0.3%
Midwest	MI	0.4%	0.3%	0.4%	0.3%
Midwest	MN	0.4%	0.3%	0.4%	0.3%
Midwest	MO	0.4%	0.3%	0.4%	0.3%
Midwest	ND	0.5%	0.4%	0.5%	0.3%
Midwest	NE	0.4%	0.3%	0.4%	0.3%
Midwest	OH	0.4%	0.3%	0.4%	0.3%
Midwest	SD	0.6%	0.3%	0.4%	0.3%
Midwest	WI	0.4%	0.3%	0.4%	0.3%
South	AL	0.3%	0.3%	0.2%	0.2%
South	AR	0.3%	0.3%	0.3%	0.2%
South	DC	0.9%	0.8%	0.8%	0.8%
South	DE	0.4%	0.3%	0.3%	0.2%
South	FL	0.4%	0.3%	0.2%	0.2%
South	GA	0.3%	0.3%	0.3%	0.2%
South	KY	0.4%	0.3%	0.3%	0.2%
South	LA	0.3%	0.3%	0.3%	0.3%
South	MD	0.4%	0.3%	0.3%	0.3%
South	MS	0.4%	0.4%	0.4%	0.4%
South	NC	0.4%	0.3%	0.3%	0.3%
South	OK	0.3%	0.3%	0.2%	0.2%
South	SC	0.4%	0.3%	0.3%	0.3%
South	TN	0.4%	0.3%	0.3%	0.2%
South	TX	0.4%	0.3%	0.3%	0.2%
South	VA	0.4%	0.3%	0.3%	0.2%
South	WV	0.4%	0.4%	0.3%	0.3%
West	AK	0.4%	0.4%	0.4%	0.4%
West	AZ	0.5%	0.2%	0.2%	0.2%
West	CA	0.4%	0.2%	0.2%	0.2%
West	CO	0.4%	0.2%	0.2%	0.2%
West	HI	0.8%	0.6%	0.7%	0.7%
West	ID	0.4%	0.3%	0.3%	0.3%
West	MT	0.5%	0.3%	0.3%	0.3%
West	NM	0.5%	0.3%	0.3%	0.3%
West	NV	0.4%	0.2%	0.2%	0.2%
West	OR	0.4%	0.3%	0.2%	0.2%
West	UT	0.5%	0.2%	0.2%	0.2%
West	WA	0.4%	0.2%	0.2%	0.2%
West	WY	0.4%	0.3%	0.3%	0.3%